センタリング理論とゼロ代名詞：日本語コーパス分析と母語話者調査の結果から

竹井　光子
広島修道大学
法学部
takeim@shudo-u.ac.jp

藤原　美保
ウィラメット大学
日本語・中国語学科
mfujiwar@willamette.edu

相沢　輝昭
広島市立大学
情報科学部
aizawa@its.hiroshima-cu.ac.jp

1　はじめに

「理論」と「実証」は常に密接な関係にある。実証に基づいた理論の構築や理論に照らした実証分析が，言語学をはじめとして様々な分野で盛んに行われている。特に，ある理論が実際に「正しいかどうか」を検証することは，研究者にとって興味深い作業である。

本稿では，Grosz らによって提唱された「センタリング理論 (Centering Theory)」の日本語データにおける妥当性を検証する。センタリング理論は，「談話の局所的な一貫性」と「話題の焦点の遷移」との関係をモデル化した理論である。日本語において焦点となりやすいのはゼロ代名詞であることから，この理論はどのような談話状況（＝焦点の遷移状態）においてゼロ代名詞の使用が好まれるのかを説明できると考えられる。そこで，ゼロ代名詞の使用に関する (1) 日本語コーパスの分析結果，と (2) 日本語母語話者の調査結果から，センタリング・モデルの妥当性の実証を試みる。

2　センタリング理論

2.1　理論の概要

センタリング理論は，談話の一貫性を構成する一側面のモデル化を試みた理論である。一貫性へのアプローチ方法としては，「連接関係」または「参照関係」に注目した 2 つの流れがあるが，センタリング理論は後者に当たる。すなわち，談話に登場する DISCOURSE ENTITY と呼ばれる名詞句が発話ごとにどのように移り変わるかに注目したアプローチである。

具体的には，前の発話（節）から引き継がれた ENTITY のうち，最も上位のものが CENTER (CB)となる。CENTER 候補となる ENTITY の順位付けは，それが文法上果たす役割を基準としている[1]。この「話題の焦点」である CENTER (CB) は発話ごとに更新され，その「更新のされ方」を TRANSITION として規定している。TRANSITION には CONTINUE (CON)，RETAIN (RET)，SHIFT (SHIFT) の 3 つがあり，その連続の仕方によって，談話の一貫性の度合，談話の解釈に必要となる推論量が異なるとしている。

Grosz らによると，CON-CON の並びは RET-RET，SHIFT-SHIFT よりも好まれる。また，RETAIN は後続の CENTER の変化を予測させる状態であり，RET-SHIFT は望ましい連続である。さらに，CENTER の移行が完了した状態から新たな CENTER が連続することは自然であり，SHIFT-CON も妥当な流れである。したがって，CON-CON は一貫性のきわめて高い連続，CON-RET-SHIFT-CON は理想的な CENTER 移行の流れであると言える。一方で，CON-SHIFT は予測なしの突然の変化，RET-CON は，変化予測を裏切る流れであり，解釈に必要とする推論量が増すと考えられる。

本稿では，理論の原典 (Grosz et al. 1983, 1986, 1995) に忠実に，TRANSITION の連続と人間が知覚する一貫性との関係を多角的に検証することにする。

2.2　分析例

前節で，Grosz らが主張する TRANSITION の連続と一貫性の度合との関係について触れたが，ここでは，談話例を (1) (2) に示して具体的に説明する。

各発話中，CENTER(CB) となっている要素を下線で示している。また，本稿では Grosz らの主張にしたがって，単独の TRANSITION ではなく，その連続に注目する。ここでは便宜上，各発話に 2 つの TRANSITION の連続を示すラベルを付すことにする。下記 (1) (2) の談話例中，このラベルを括弧内太字で示し，以後これを「連続パターン」と呼ぶことにする。例えば，**(CON-SHIFT)** ラベルの連続パターンは CONTINUE の発話に続く SHIFT の発話であることを，**(RET-SHIFT)** は RETAIN の発話に続く SHIFT の発話であることを示す。TRANSITION の連続を見るのは，同じ SHIFT の発話であっても，何の発話の後続であるかが一貫性の度合を測る上で重要だからである。

[1] 本稿では，主格［が］＞ 主格以外（対格［を］，与格［に］，連体格［の］，場所格［で］など）の順とする。本稿で採用した定義や分析基準の詳細については，Yamura-Takei (2005)にある。

談話例 (1)

1. 江戸時代には藩がありました。

2. 藩は今の県とだいたい同じです。
   CON
3. 藩に大名がいました。
   RET **(CON – RET)**
4. 大名は自分の藩と江戸にうちがありました。
   SHIFT **(RET – SHIFT)**
5. 大名は藩に一年，江戸に 1 年住まなければなりませんでした。
   CON **(SHIFT – CON)**
6. 大名の妻と子どもは江戸に住んでいました。
   RET **(CON – RET)**
7. 大名は 1 年おきに江戸まで歩いて行かなければなりませんでした。
   CON **(RET – CON)**

(1-1)で導入された ENTITY のうち，(1-2)で「藩」が CENTER となるが (CON)，(1-3)で主格以外の位置につくことで CENTER の変化を予測させ (RET)，(1-4)で新たに主格の位置についた「大名」へと CENTER が移行する (SHIFT)。さらに(1-5)では，その新 CENTER が継続している (CON)。すなわち，(1-2) から(1-5) は，CON-RET-SHIFT-CON の円滑な流れとなっている。

次の(1-6)では，CENTER であった「大名」を連体格に置き，あらたな ENTITY である「妻と子ども」を導入することで，再び CENTER の変化を予測させているが (RET)，(1-7)ではその予測に反して「大名」に CENTER が戻っている (CON)。これは，(1-3)～(1-4)の円滑な流れ (CON-RET-SHIFT)と異なり，焦点の頻繁な変化による予想外の流れであり，知覚する一貫性の度合が低いと考えられる (CON-RET-CON)。

談話例 (2)

1. 江戸時代の人はみんなお寺に名前を登録しなければなりませんでした。

2. そして江戸時代の人は牛肉を食べてはいけませんでした。
   CON **(CON – CON)**
3. 牛肉は畑しごとを手伝う大切な動物の肉だからです。
   SHIFT **(CON - SHIFT)**

(2-1) の CENTER である「大名」が (2-2)でも連続して用いられているが (CON)，(2-3)では前発話で下位（対格）にあった ENTITY「牛肉」が突然 CENTER に抜擢されている (SHIFT)。この CON-SHIFT の流れも，焦点の突然の変化による唐突さから一貫性の度合は低くなる。

2.3 ゼロ代名詞とセンタリング

談話における一貫性と推論量には，指示表現が大いに関係する。英語においては，名詞句を用いるか代名詞を用いるかが推論量に影響を与える。日本語において，この代名詞に相当するのがゼロ代名詞である。

ゼロ代名詞とは，文脈などから推測が可能であると判断できる場合に省略という形をとる項（連用項および連体項）のことである。この判断を日本語母語話者は直感的に行っている。しかし，この直感的判断の裏には何らかのメカニズムが存在するはずである。冗長さのために不自然にならずかつ省略によって曖昧さが生じたり多大な推論量を課したりすることのないようバランスをとりながら判断を行う。その判断を制御するメカニズムの一つと考えられるのがセンタリングであると考える。

すなわち，一貫性が高く多くの推論量を必要としない談話状況においては省略が行われるであろうし，逆に一貫性が低い状況においては省略をさけるであろうということが仮定できる。

この仮定を統計的に実証するため，前述の 8 つの TRANSITION の連続パターン **(CON-CON, RET-RET, SHIFT-SHIFT, CON-RET, RET-SHIFT, SHIFT-CON, CON-SHIFT, RET-CON)** において，(1) ゼロ代名詞が CENTER となっている割合を求めるコーパス分析，(2) CENTER をゼロ代名詞化してもよいかについての母語話者の直感的判断，の 2 点を調査した。

3 コーパス分析

3.1 分析方法

コーパスは，8 つの日本語教科書中の読解教材 83 テキスト（2,007 節）を使用し，人手によりゼロ代名詞の検出，CENTER の特定，TRANSITION の計算を行った。

分析の目的は，TRANSITION の連続パターンごとに CENTER がゼロ代名詞化されている割合を算出することである。一貫性の高い（多くの推論量を要求しない）談話状況においてはゼロ代名詞が使用されるはずであるとの仮定にもとづき，8 つの TRANSITION の連続パターンの一貫性の高さによる順位付けを試みる。

3.2 分析結果

コーパス 2,007 発話（節）分に含まれる CENTER の総数は 1,248 であった。これは，談話頭など，前発話から引き継ぐ CENTER が存在しない場合があるからである。そのうち 841(67%) がゼロ代名詞であった。

図 1，2 に TRANSITION の連続パターンごとにその数と CENTER がゼロ代名詞（ZERO）であるか名詞表現（non-ZERO）であるかの割合を示す。

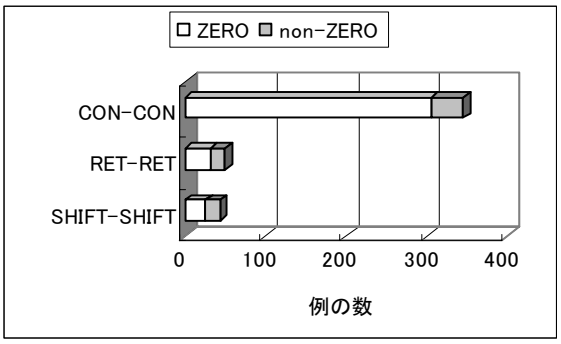

図 1：コーパス分析結果（1）

Grosz らが最も好ましいとする **CON-CON** は圧倒的に数も多く，CENTER (CB) がゼロ代名詞である割合も 88%と高い。一方，**RET-RET**，**SHIFT-SHIFT** は数も少なく，それぞれ 62%，58%とゼロ代名詞の割合が下がる。

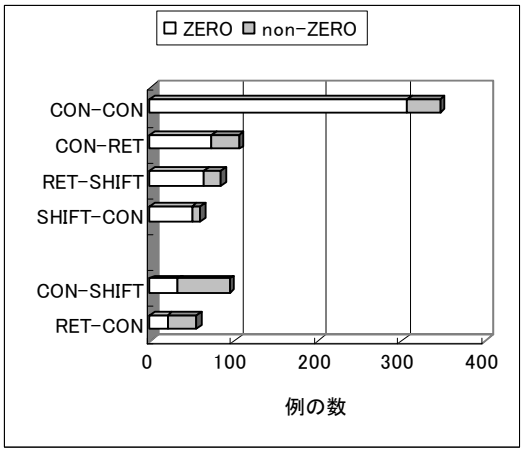

図 2：コーパス分析結果（2）

また，Grosz らが妥当な流れだとする 4 つの TRANSITION 連続であるが，前述の **CON-CON** の 88%に続き，**CON-RET** では 60%，**RET-SHIFT** では 76%，**SHIFT-CON** では 83%の発話において CENTER をゼロ代名詞で実現している。**CON-RET**，**RET-SHIFT** において割合が比較的低いのは，CENTER の移行過程にあることから曖昧さを避けるためにゼロ代名詞の使用が控えられていると考えられる。

一方で，逸脱した流れである **CON-SHIFT** では 33%，**RET-CON** では 38%と，明らかにゼロ代名詞の割合が低くなっている。

実際のテキストでゼロ代名詞が使われている割合が一貫性（および解釈に必要な推論量）を示す指標となると仮定して分析を行ったが，明らかな傾向の違いが結果として出た点は興味深い。また，Grosz らの主張とも一致する結果となった。

4 母語話者調査

4.1 調査方法

日本語母語話者（大学教員および学生）50 名に対して，アンケート調査を行った。被験者には，テキスト（5 段落，18 文）を読み，「より自然な日本語にするため」という観点から，下線が引かれた CENTER に当たる名詞句を明示したままの方がよいか，省略した方がよいかについて判断を求めた。前述の談話例 (1) (2) は，調査に用いたサンプルの一部である。

ここでも，ゼロ代名詞の使用を認める割合が高いほど，その談話状況（TRANSITION の連続パターン）の一貫性が高い（よって曖昧さの発生などの危険が生じない）と母語話者が判断したと仮定する。

4.2 調査結果

図 3 に，各 TRANSITION の連続パターンにおいてゼロ代名詞を使用する方(ZERO)・しない方(non-ZERO)がよいと判断した率を示す。一連続パターンにつき複数のサンプルがある場合には平均値を用いている。

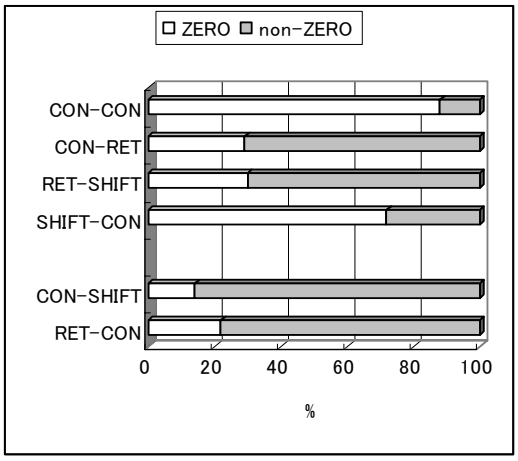

図 3：母語話者調査結果

最も一貫性が高いとされる **CON-CON** では，88%の被験者がゼロ代名詞の使用を認めている。また，CENTER 移行後の連続である **SHIFT-CON** においても 72%という比較的高い割合でゼロ代名詞の使用を認めている。**CON-RET** が 29%，**RET-SHIFT** が 30%と低いのはコーパス分析結果と同様，CENTER の移行過程にあることが影響していると思われる。

一方で，唐突な流れである **CON-SHIFT** では 14%，**RET-CON** では 22%と，明らかにゼロ代名詞を許容する割合が低くなっている。例えば，前述の **RET-CON**

の発話例 (1-7) において，CENTER の「大名は」をゼロ代名詞化するとする。この前の発話 (1-6) は，「大名」を連体格に下げ，新しく「妻と子ども」を導入した RETAIN の状態である。センタリングの観点からは「妻と子ども」への話題の転換を期待させる。しかし，ここではその期待を裏切り「大名」が再び CENTER となっており，ゼロ代名詞が指示するものが「大名」であることを正しく解釈するには，「1 年おきに江戸に歩いて行く」のは「妻や子ども」ではなく「大名」自身であるという背景知識からの推論が必要となる。ゼロ代名詞化を認めない意見が多いのは，この推論要求から安全策をとった結果と思われる。

このように，母語話者の判断に一貫性が高い場合にはゼロ代名詞を使用し，低い場合には使用しないという明らかな傾向の違いが出たことは興味深い。母語話者は，センタリングが説明する一貫性の度合の差を直感的に感じとって省略に関する判断を下しているのではないだろうか。

5 まとめ

センタリング理論の主張の妥当性を，コーパス分析と母語話者調査の 2 点から検証した。TRANSITION の連続パターンごとに，CENTER がゼロ代名詞となっている割合（コーパス分析）・CENTER のゼロ代名詞化を支持する割合（母語話者調査）を表 1 にまとめた。割合の高い順としている。[2]

表 1：連続パターンとゼロ代名詞の使用

	コーパス	母語話者
CON-CON	88%	88%
SHIFT-CON	83%	72%
RET-SHIFT	76%	30%
RET-RET	62%	-
CON-RET	60%	29%
SHIFT-SHIFT	58%	-
RET-CON	38%	22%
CON-SHIFT	35%	14%

コーパス分析の数値と母語話者調査の数値が傾向的に一致している[3]。このゼロ代名詞の割合が，それを含む談話状況の一貫性の度合・その談話を解釈するために必要とする推論の量を示す指標となるとの仮定のもとに調査を行ったが，Grosz らの主張とも一致する結果となった。

最も一貫性が高いとされる **CON-CON**，自然な流れである **SHIFT-CON** から，唐突な流れである **RET-CON**, **CON-SHIFT** まで，一貫性の度合により段階的な序列ができている。このことから，センタリング理論の妥当性が日本語データにおいて証明されたと言えるだろう。

中間的な一貫性を持つ連続パターン（表の網掛け部分）においては，平行構造，接続関係，背景知識などセンタリング以外の推論材料がゼロ代名詞使用か否かの判断に影響を与えていると思われる。その詳細な分析は今後の課題の一つである。

センタリング理論は計算機的談話理論であり，その柔軟さゆえに様々な改変，拡張，応用が行われてきた。初期に BFP アルゴリズム(Brennan et al, 1987) が提案されたため，照応解析のための理論であるという印象が強い。しかし，これはこの理論が元来意図したことではなく，むしろ Kameyama (1985) が指摘するように，そのメカニズムを「人間の言語能力 (linguistic competence)」の一部であり「談話処理における認知過程モデル」と見るべきであろう。

省略についての直感的判断を説明することは難しい。本調査の結果が，自然言語処理に限らず，日本語教育の場でも活用されることを願っている。「どこで何を省略すべきか，省略しても安全なのか」についての直感的認知過程に，実証的データに基づく理論的説明を与えることで，省略を指導する日本語教員の一助になればと考える。

参考文献

Brennan, Suzan E., Marilyn Walker Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford, CA*, 155-162.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1983. Providing a unified account of noun phrase in discourse. In *Proceedings of the 21th Annual Meeting of the Association for Computational Linguistics, Cambridge, MA*, 44-50.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1986. Towards a computational theory of discourse interpretation. Unpublished manuscript.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21/2, 203-225.

Kameyama, Megumi. 1985. *Zero Anaphora: The Case of Japanese*. PhD dissertation, Stanford University.

Yamura-Takei, Mitsuko. 2005. *Theoretical, Technological and Pedagogical Approaches to Zero Arguments in Japanese Discourse: Making the Invisible Visible*. Doctoral thesis, Hiroshima City University.

[2] RET-RET, SHIFT-SHIFT については，母語話者調査用のサンプルに含まれていなかったため空欄となっている。

[3] 母語話者調査結果の数値が全般的に低いのは，明示化された CENTER を与え，ゼロ代名詞化した方が自然かどうかを問う質問形式であったことが影響していると思われる。